# A generalization of the Erdős-Turán law for the order of random permutation

Alexander Gnedin[*], Alexander Iksanov[†] and Alexander Marynych[‡]

January 20, 2013

**Abstract**

We consider random permutations derived by sampling from stick-breaking partitions of the unit interval. The cycle structure of such a permutation can be associated with the path of a decreasing Markov chain on $n$ integers. Under certain assumptions on the stick-breaking factor we prove a central limit theorem for the logarithm of the order of the permutation, thus extending the classical Erdős-Turán law for the uniform permutations and its generalization for Ewens' permutations associated with sampling from the PD/GEM($\theta$) distribution [3]. Our approach is based on using perturbed random walks to obtain the limit laws for the sum of logarithms of the cycle lengths.

Keywords: random permutation, Erdős-Turán law, stick-breaking, perturbed random walk

## 1 Introduction

Let $\mathfrak{S}_n$ be the symmetric group on $[n] := \{1, \dots, n\}$. The order of permutation $\sigma \in \mathfrak{S}_n$ is the smallest positive integer $k$ such that the $k$-fold composition of $\sigma$ with itself is the identity permutation. The order can be determined from the cycle representation of $\sigma$ as the least common multiple (l.c.m.) of the cycle lengths. For instance, permutation $\sigma = (1\,9\,6\,2)(3\,7\,5)(4\,8)$ has order 12.

A random permutation $\Pi_n$ of $[n]$ is a random variable with values in the set $\mathfrak{S}_n$. A widely known parametric family of random permutations has probability mass function

$$\mathbb{P}\{\Pi_n = \sigma\} = c^{-1}\,\theta^{|\sigma|}, \qquad \theta > 0, \tag{1}$$

where $|\sigma|$ denotes the number of cycles, and the constant is $c = (\theta)_n := \Gamma(\theta + n)/\Gamma(\theta)$. This family is sometimes called Ewens' permutations since the collection of cycle lengths is then a random partition distributed according to the Ewens sampling formula [3, 29].

[*] Queen Mary University of London, e-mail: a.gnedin@qmul.ac.uk

[†] National Taras Shevchenko University of Kiev, e-mail: iksan@univ.kiev.ua

[‡] National Taras Shevchenko University of Kiev, e-mail: marynych@unicyb.kiev.ua

The instance $\theta = 1$ corresponds to the uniform distribution under which all permutations $\sigma \in \mathfrak{S}_n$ are equally likely.

For random permutation $\Pi_n$ with some fixed distribution let $K_{n,r}$ be the number of cycles of length $r$ and let $K_n := |\Pi_n| = \sum_{r=1}^n K_{n,r}$ be the total number of cycles. We call vector $(K_{n,1}, \dots, K_{n,n})$ the *cycle partition* of $\Pi_n$. In terms of the cycle partition the order of $\Pi_n$ is the random variable defined as

$$O_n := \text{l.c.m.}\{r \in [n] : K_{n,r} > 0\}. \tag{2}$$

In a seminal 1967 paper [9] Erdős and Turán showed that for the uniform permutation the distribution of $\log O_n$ is asymptotically normal. Arratia and Tavaré [4] extended this result to Ewens' permutations, by showing that

$$\frac{\log O_n - (\theta/2)\log^2 n}{\sqrt{(\theta/3)\log^3 n}} \overset{d}{\to} \mathcal{N}(0,1), \qquad n \to \infty. \tag{3}$$

The proof in [4] (see also [3], Theorem 5.15), apparently the shortest one known, is based on the Feller coupling and asymptotic independence of the $K_{n,r}$'s.

In this paper we generalize the Erdős-Turán law to a much richer family of random permutations derived from stick-breaking partitions of the unit interval by means of a simple occupancy scheme called Kingman's 'paintbox process' [29]. A toolbox of methods suitable for the study of Ewens' permutations is no longer applicable in the wider setting due to the lack of asymptotic independence of the $K_{n,r}$'s. Instead, extending the line initiated in [10, 14, 15, 16, 17, 20, 21], we apply the methods of renewal theory to obtain results on the weak convergence of the decisive quantity $\log T_n := \sum_r K_{n,r} \log r$ which approximates the logarithm of the order of permutation. We show that the normal and other stable distributions can appear as limit laws, as determined by properties of the stick-breaking factor.

There have been many studies of random permutations that are conditionally uniform given the value of some permutation statistic [8, 12, 18, 6]. Our motivation to consider the class of stick-breaking models has several sources, among which are the theory of regenerative composition structures [19], more general exchangeable partitions [29] and the logarithmic combinatorial structures [3]. The present paper is the first study of a *separable statistic* $\sum_r K_{n,r} h(r)$ with unbounded function $h$ for the partitions of integers derived from the general stick-breaking. It would be interesting to further study separable statistics and approximations to $O_n$ for other permutation models associated with exchangeable partitions.

The organization of the rest of the paper is as follows. In Section 2 we introduce the class of permutations derived from the stick-breaking. The principal results are formulated in Section 3. In Section 4 we prove that under various regularity conditions $\log T_n$ yields a good approximation to $\log O_n$ with an error term of the order $o(\log^{3/2} n)$. In Section 5 we investigate the weak convergence of $\log T_n$ and prove Theorem 3.2; the method here exploits a link between the $K_{n,r}$'s and certain perturbed random walks. Theorem 3.1 which is our generalization of the Erdős-Turán law follows then as a corollary. The auxiliary results used in the proofs are collected in the Appendix.

# 2 Permutations derived from stick-breaking

**The Basic Construction** Let $W$ be a random variable, called *stick-breaking factor*, with values in $(0,1)$. Consider a multiplicative renewal point process $\mathcal{Q}$ with atoms

$$Q_0 := 1, \quad Q_j := \prod_{i=1}^{j} W_i, \quad j \in \mathbb{N},$$

where $W_i$ are independent replicas of $W$. The gaps in $\mathcal{Q}$ yield a partition of $[0,1]$ in infinitely many intervals $(Q_{j+1}, Q_j]$ accumulating near 0. Let $U_1, \dots, U_n$ be a sample from the uniform $[0,1]$ distribution, independent of $\mathcal{Q}$. A random permutation $\Pi_n$ is defined by organizing integers $i_1, \dots, i_\ell$ in a cycle $(i_1 \ \dots \ i_\ell)$ if the following occur:

(i) $U_{i_1} < \cdots < U_{i_\ell}$,

(ii) the sample points $U_{i_1}, \dots, U_{i_\ell}$ fall in the same interval $(Q_{j+1}, Q_j]$ ,

(iii) only $U_{i_1}, \dots, U_{i_\ell}$ out of $U_1, \dots, U_n$ fall in this interval $(Q_{j+1}, Q_j]$.

Listing the sample points in increasing order and inserting a $|$ between two neighbouring order statistics if they belong to distinct component intervals of $(0,1] \setminus \mathcal{Q}$, the cycle notation of $\Pi_n$ is read left-to-right.

For instance, the list $U_7 \,|\, U_3 \ U_4 \ U_2 \ U_5 \,|\, U_6 \ U_1$ yields permutation $(7)(3\ 4\ 2\ 5)(6\ 1)$. To pass to the standard cycle notation $(1\ 6)(2\ 5\ 3\ 4)(7)$ one needs to re-arrange the cycles in the order of increase of their minimal elements, and to rotate each cycle so that the least element of the cycle appears first. We prefer, however, to write the cycles and the elements within the cycles in accord with the natural order on reals, as dictated by the Basic Construction. A reason for this ordering of cycles is the following recurrence property:

- *Regeneration*: for $m \in \{1, \dots, n-1\}$, conditionally given the last cycle of $\Pi_n$ has length $m$, the cycle partition of $\Pi_n$ with the last cycle deleted has the same distribution as the cycle partition of $\Pi_{n-m}$.

It is straightforward from the construction that $\Pi_n$ also satisfies:

- *Coherence*: permutations $\Pi_n$ are defined consistently for all values of $n$. Passing from $\Pi_{n+1}$ to $\Pi_n$ amounts to removing integer $n+1$ from a cycle.

- *Exchangeability*: the distribution of $\Pi_n$ is invariant under conjugations in $\mathfrak{S}_n$. Equivalently, given the cycle partition $(K_{n,1}, \dots, K_{n,n})$ the distribution of $\Pi_n$ is uniform.

In combination with exchangeability, the regeneration property can be re-stated as follows: given the last cycle of $\Pi_n$ is of length $m$, a permutation resulting from deletion of the last cycle and re-labeling the remaining elements by the increasing bijection with $[n-m]$ is a distributional copy of $\Pi_{n-m}$.

There are two further useful ways to generate the cycle partition of $\Pi_n$.

**A Markov chain representation** Consider a decreasing Markov chain on nonnegative integers with absorption at 0 and the *decrement matrix*

$$q(n,m) = \binom{n}{m} \frac{\mathbb{E}[W^{n-m}(1-W)^m]}{1-\mathbb{E}W^n}, \qquad 1 \le m \le n, \tag{4}$$

specifying transition probabilities from $n$ to $n-m$. For the Markov chain $M_n$ starting at $n$, $K_{n,r}$ is the number of jumps of size $r$ on the path of $M_n$ from $n$ to 0. The arrangement of the cycle lengths in the Basic Construction corresponds to the decrements of $M_n$ written in the time-reversed order.

**The infinite occupancy scheme** This model is sometimes called the *Bernoulli sieve* [10, 14, 15, 16, 17, 20, 21]. Think of the gaps $(Q_j, Q_{j-1}]$ as boxes $1, 2, \dots$ with frequencies

$$P_j := W_1 W_2 \cdots W_{j-1}(1-W_j), \quad j \in \mathbb{N}. \tag{5}$$

Given the frequencies, balls $1, 2, \dots$ are thrown independently so that each ball hits box $j$ with probability $P_j$. Then $K_{n,r}$ is the number of boxes occupied by exactly $r$ out of the first $n$ balls.

**Additive renewal process representation** Mapping $(0,1]$ to $\mathbb{R}_+$ via $x \mapsto -\log x$ sends $\mathcal{Q}$ to the additive renewal process with the generic increment $-\log W$, and sends the uniform sample to a sample from the standard exponential distribution. The construction of permutation and the occupancy scheme are obviously re-stated in the new variables.

It has been observed (see [11], Theorem 2.1) that the instance of Ewens' permutation fits in the Basic Construction by choosing a factor $W \overset{d}{=} \text{beta}(\theta, 1)$, with the density

$$\mathbb{P}\{W \in \mathrm{d}x\} = \theta x^{\theta-1}\mathrm{d}x, \qquad x \in (0,1).$$

A better known connection of Ewens' $\Pi_n$ to the stick-breaking stems from the fact that the scaled by $n$ lengths of the cycles in the normalized notation converge as $n \to \infty$ to $(P_1, P_2, \dots)$ as in (5) with $W_j \overset{d}{=} \text{beta}(\theta, 1)$. The distribution of the limit is known as the GEM($\theta$) law, which is related to the Poisson-Dirichlet PD($\theta$)-distribution through a size-biased permutation of the terms. As a finite-$n$ counterpart of this dual role of the stick-breaking, the sequence of lengths of cycles ordered by increase of the minimal elements and the reversed sequence of the cycle lengths derived from the Basic Construction have the same distribution. In particular, both sequences can be identified with the sequence of decrements of the Markov chain $M_n$ with decrement matrix

$$q(n, n-m) = \binom{n}{m} \frac{(\theta)_{n-m} m!}{(\theta+1)_{n-1} n}.$$

It follows from a result of Kingman that the coincidence of distributions of the two different arrangements of the unordered set of the cycle-lengths characterizes the Ewens permutation within the family of random permutations with the regenerative property, see [13] for this fact and variations.

We note in passing that by a version of the Basic Construction each system of coherent random permutations $(\Pi_n)_{n\in\mathbb{N}}$ with the properties of exchangeability and regeneration, with respect to deletion of a cycle of $\Pi_n$ chosen by some random rule, uniquely corresponds to a random regenerative subset of $\mathbb{R}_+$ which coincides with the closed range of a subordinator $S$ [19] . Distinguishing features of the subfamily in focus in the present paper are: (1) $S$ is a compound Poisson process with jumps distributed like $|\log W|$; (2) the last cycle of $\Pi_n$ has the length of the order $O(n)$ as $n$ grows.

# 3 Main results

In the sequel we use the following notation for the moments of the stick-breaking factor

$$\mu := \mathbb{E}|\log W|, \quad \sigma^2 := \mathrm{Var}\,(\log W) \quad \text{and} \quad \nu := \mathbb{E}|\log(1-W)|,$$

which may be finite or infinite. We shall also use the notion of slow variation. Function $\ell : (0,\infty) \to (0,\infty)$ is called *slowly varying at* $\infty$ if for all $\lambda > 0$,

$$\lim_{x\to\infty} \frac{\ell(\lambda x)}{\ell(x)} = 1.$$

Our purpose is to extend (3) to a wider class of random permutations $\Pi_n$ derived from stick-breaking, along the following lines.

**Theorem 3.1.** *Suppose the law of $W$ is absolutely continuous with a density $f$.*

I. *If there exist $\delta_1 \geq 0$ and $\delta_2 \geq 0$ such that $f$ is nonincreasing on $(0,\delta_1)$, bounded on $[\delta_1, 1-\delta_2]$ and nondecreasing on $(1-\delta_2, 1)$ then*

   (a) *If $\sigma^2 < \infty$ then, with*

   $$b_n = \mu^{-1}\Big(2^{-1}\log^2 n - \int_0^{\log n}\int_0^z \mathbb{P}\{|\log(1-W)| > x\}\mathrm{d}x\mathrm{d}z\Big) \qquad (6)$$

   *and $a_n = ((3\mu^3)^{-1}\sigma^2\log^3 n)^{1/2}$, the limiting distribution of $(\log O_n - b_n)/a_n$ is standard normal.*

   (b) *If $\sigma^2 = \infty$, and*

   $$\int_0^x y^2\,\mathbb{P}\{|\log W| \in \mathrm{d}y\} \;\sim\; \ell(x), \quad x \to \infty,$$

   *for some $\ell$ slowly varying at $\infty$, then, with $b_n$ given by (6) and*

   $$a_n = (3\mu^3)^{-1/2}c_{[\log n]}\log n,$$

   *where $(c_n)$ is any positive sequence satisfying $\lim_{n\to\infty} n\ell(c_n)/c_n^2 = 1$, the limiting distribution of $(\log O_n - b_n)/a_n$ is standard normal.*

(c) *If*

$$\mathbb{P}\{|\log W| > x\} \ \sim\ x^{-\alpha}\ell(x), \quad x \to \infty, \tag{7}$$

*for some $\ell$ slowly varying at $\infty$ and $\alpha \in (1,2)$ then, with $b_n$ as in* (6) *and*

$$a_n = ((\alpha+1)\mu^{\alpha+1})^{-1/\alpha} c_{\lfloor \log n \rfloor} \log n,$$

*where $(c_n)$ is any positive sequence satisfying $\lim_{n\to\infty} n\ell(c_n)/c_n^\alpha = 1$, the limiting distribution of $(\log O_n - b_n)/a_n$ is the $\alpha$-stable law with characteristic function*

$$u \mapsto \exp\{-|u|^\alpha \Gamma(1-\alpha)(\cos(\pi\alpha/2) + i\sin(\pi\alpha/2)\,\mathrm{sgn}(u))\}, \ u \in \mathbb{R}. \tag{8}$$

II. *If for some $\alpha \in [0,1)$*

$$\sup_{x\in[0,1]} x^\alpha(1-x)^\alpha f(x) < \infty; \tag{9}$$

*then $\sigma^2 < \infty$ and*

$$\frac{\log O_n - (2\mu)^{-1}\log^2 n}{\sqrt{(3\mu^3)^{-1}\sigma^2 \log^3 n}} \ \overset{d}{\to}\ \mathcal{N}(0,1), \quad n \to \infty.$$

In particular, these conditions cover all bounded densities, and all beta$(a,b)$ densities with arbitrary parameters $a, b > 0$. Following an approach exploited by previous authors we derive our extension of the Erdős-Turán law in two steps. We first show that the accompanying quantity $\log T_n$ yields a good approximation to $\log O_n$, where

$$T_n := \prod_{r=1}^n r^{K_{n,r}}, \tag{10}$$

is the product of cycle lengths of $\Pi_n$. Then we study the weak convergence of $\log T_n$.

Functional $\log T_n$ is an instance of a *separable statistic* of the form $\sum_r K_{n,r}h(r)$ (the terminology is borrowed from [26, 27], where it was used in the context of occupancy problems). Functionals $K_{n,r}$ and $K_n$ are themselves of this kind with some indicator functions $h$, but for $\log T_n$ the function $h$ is unbounded. For Ewens' permutations quite general separable statistics were studied by Babu and Manstavičius, see e.g. [5, 25].

**Theorem 3.2.** *If $W$ satisfies the moment conditions required, respectively, in parts* (a), (b) *and* (c) *of* Theorem 3.1, *then the conclusions of parts* (a), (b) *and* (c) *hold with* $\log O_n$ *replaced by* $\log T_n$, *without the assumption regarding the existence of density of $W$.*

**Example: beta distributions** Assuming $W \overset{d}{=} \text{beta}(\theta, 1)$ we have $\mu = \theta^{-1}$, $\sigma^2 = \theta^{-2}$ and

$$\lim_{n\to\infty} \frac{\int_0^{\log n}\int_0^z \mathbb{P}\{|\log(1-W)| > x\}\mathrm{d}x\mathrm{d}z}{\log^{3/2} n} = 0,$$

since the numerator is $O(\log n)$. Application of Theorem 3.2 (a) yields

$$\frac{\log T_n - (\theta/2)\log^2 n}{\sqrt{(\theta/3)\log^3 n}} \ \overset{d}{\to}\ \mathcal{N}(0,1), \quad n \to \infty,$$

which was previously obtained in [4], equation (34).

# 4 Approximation of $\log O_n$ by $\log T_n$

For $j \in [n]$ set

$$D_{n,j} := \sum_{r \le n,\, j|r} K_{n,r} = \sum_{r=1}^{\lfloor n/j \rfloor} K_{n,rj}.$$

For the later use we need appropriate bounds for the expectation $\mathbb{E}(D_{n,j}-1)^+$.

**Lemma 4.1.** *Under the assumptions of* Theorem 3.1 *the asymptotic relations*

$$\mathbb{E}(D_{n,j}-1)^+ = O\left(\frac{\log n}{j}\right), \tag{11}$$

$$\mathbb{E}(D_{n,j}-1)^+ = O\left(\frac{\log^2 n}{j^2}\right) \tag{12}$$

*hold uniformly in $j \in [n]$.*

*Proof.* Define $D^{(1)}_{n,j} := \sum_{r=1}^{\lfloor n/j \rfloor -1} K_{n,rj}$. It is obvious that $(D_{n,j}-1)^+ \le D^{(1)}_{n,j}$.

Let $A_n$ be the length of the last cycle of $\Pi_n$, with distribution $\mathbb{P}\{A_n = j\} = q(n,j)$ as in (4). One can check that the bivariate array $(D^{(1)}_{n,j})$ satisfies the distributional recurrence

$$\begin{aligned} D^{(1)}_{n,j} &= 0, \;\; n<j, \\ D^{(1)}_{n,j} &\overset{d}{=} 1_{\{j|A_n,\, j \le A_n \le n-j\}} + \hat{D}^{(1)}_{n-A_n,j}, \;\; n \ge j, \end{aligned} \tag{13}$$

where the variables $\hat{D}^{(1)}_{n,k}$ are assumed independent of $\Pi_n$ and marginally distributed like $D^{(1)}_{n,k}$ for all $n,k \in \mathbb{N}$. Taking expectations yields

$$\mathbb{E}D^{(1)}_{n,j} = \sum_{r=1}^{\lfloor n/j \rfloor -1} \mathbb{P}\{A_n = rj\} + \sum_{i=j}^{n} \mathbb{P}\{n - A_n = i\}\mathbb{E}D^{(1)}_{i,j} \quad \text{for } n \ge j,$$

and $\mathbb{E}D^{(1)}_{n,j} = 0$ for $n < j$.

By Lemma 6.2

$$j \sum_{r=1}^{\lfloor n/j \rfloor -1} \mathbb{P}\{A_n = rj\} = O(1), \;\; j \le n, \;\; n \in \mathbb{N}. \tag{14}$$

Now relation (11) follows by the virtue of part (i) of Lemma 6.1 and Lemma 6.3 with $c_j = j$.

To prove the second assertion (12), note that

$$\begin{aligned} (D_{n,j}-1)^+ &= (D_{n,j}-1)^+ 1_{\{K_{n,\lfloor n/j \rfloor j}=0\}} = (D^{(1)}_{n,j}-1)^+ 1_{\{K_{n,\lfloor n/j \rfloor j}=0\}} \\ &\le (D^{(1)}_{n,j}-1)^+ \le D^{(1)}_{n,j}(D^{(1)}_{n,j}-1)/2 =: D^{(2)}_{n,j}, \end{aligned}$$

holds almost surely. Squaring relation (13) and using (14) and (11) yield

$$\mathbb{E}D_{n,j}^{(2)} = O(j^{-2}\log n) + \sum_{i=j}^{n} \mathbb{P}\{n - A_n = i\}\mathbb{E}D_{i,j}^{(2)}, \quad n \geq j, \ j \in \mathbb{N},$$

Finally, application of part (ii) of Lemma 6.1 and Lemma 6.3 with $c_j = j^2$ establish (12), as wanted. □

The following estimate of the difference $\log T_n - \log O_n$ generalizes Lemma 4 in [4].

**Lemma 4.2.** *Under the assumptions of* Theorem 3.1 *the following asymptotic relations hold*

$$\mathbb{E}\left(\log T_n - \log O_n\right) = O\left(\log n \log\log n\right), \quad n \to \infty.$$

*Proof.* We start with a known representation (p. 289 in [24])

$$\log T_n - \log O_n = \sum_{p\in\mathcal{P}} \log p \sum_{s\geq 1} (D_{n,\,p^s} - 1)^+,$$

where $\mathcal{P}$ denotes the set of prime numbers, which implies

$$\begin{aligned}
\mathbb{E}\left(\log T_n - \log O_n\right) &= \sum_{p\in\mathcal{P},\, s\geq 1} \log p\, \mathbb{E}(D_{n,\,p^s} - 1)^+ \\
&\leq \sum_{p\in\mathcal{P},\, p\leq \log n} \log p\, \mathbb{E}(D_{n,p} - 1)^+ + \sum_{p\in\mathcal{P},\, s\geq 2,\, p^s\leq \log n} \log p\, \mathbb{E}(D_{n,p^s} - 1)^+ \\
&+ \sum_{j>\log n} \log j\, \mathbb{E}(D_{n,j} - 1)^+ =: S_1(n) + S_2(n) + S_3(n).
\end{aligned}$$

Applying (11) along with Theorem 4.10 in [2] which states that

$$\sum_{p\in\mathcal{P},\, p\leq x} \frac{\log p}{p} = \log x + O(1), \quad x \to \infty,$$

proves $S_1(n) = O(\log n \log\log n)$. Using (11) again yields $S_2(n) = O(\log n)$. Finally, from (12) we infer $S_3(n) = O(\log n \log\log n)$. The proof is complete. □

# 5 Weak convergence of $\log T_n$

To prove Theorem 3.2 we shall exploit a strategy as in [14] (see also [20]), which amounts to connecting the asymptotics of $\log T_n$ (as $n \to \infty$) with that of the 'small frequencies' $P_k$ (as $k \to \infty$). Since the process $(\log P_k)_{k\in\mathbb{N}}$ defined by (5) is a particular *perturbed random walk*, we start in Subsection 5.1 with developing necessary backgrounds on the perturbed random walks. These results are further specialized to $\log P_k$ in Subsection 5.2, which eventually allows to complete the proof of Theorem 3.2.

### 5.1 Results for perturbed random walks

Let $(\xi_k, \eta_k)_{k\in\mathbb{N}}$ be independent copies of a random vector $(\xi, \eta)$ with arbitrarily dependent components $\xi > 0$ and $\eta \geq 0$. We assume that the law of $\xi$ is nondegenerate and that the law of $\eta$ is not the Dirac mass at 0. Set $F(x) := \mathbb{P}\{\eta \leq x\}$ and $r(x) := \int_0^x (1 - F(y))\mathrm{d}y$.

For $(S_k)_{k\in\mathbb{N}_0}$ a random walk with $S_0 = 0$ and increments $\xi_k$, the sequence $(T_k)_{k\in\mathbb{N}}$ with

$$T_k := S_{k-1} + \eta_k, \quad k \in \mathbb{N},$$

is called a *perturbed random walk*. Since $\lim\limits_{k\to\infty} T_k = \infty$ a.s., there is some finite number

$$N(x) := \#\{k \in \mathbb{N} : T_k \leq x\}, \quad x \geq 0,$$

of sites visited on the interval $[0, x]$. Set also

$$\rho(x) := \#\{k \in \mathbb{N}_0 : S_k \leq x\} \;=\; \inf\{k \in \mathbb{N} : S_k > x\}, \quad x \geq 0,$$

and

$$M(x) := \sum_{k\geq 0} \mathbb{E}\left(1_{\{T_{k+1}\leq x\}} \middle| S_k\right) = \sum_{k\geq 0} F(x - S_k), \quad x \geq 0.$$

The main result of this subsection is given next.

**Theorem 5.1.** *Assume that* $\mathtt{m} := \mathbb{E}\xi < \infty$ *and*

$$\frac{\rho(x) - \mathtt{m}^{-1}x}{c(x)} \overset{d}{\to} Z, \quad x \to \infty.$$

*Then*

$$I(x) := \frac{\int_0^x (N(y) - \mathtt{m}^{-1}(y - r(y)))\mathrm{d}y}{xc(x)} \overset{d}{\to} \int_0^1 Z(y)\mathrm{d}y =: X, \quad x \to \infty,$$

*where* $(Z(t))_{t\geq 0}$ *is a stable Lévy process such that* $Z(1)$ *has the same law as* $Z$.

*Remark* 5.2. It is known (see Proposition 27 in [28]) that $c(x) \;\sim\; x^\beta \ell_1(x)$ for some $\beta \in [1/2, 1)$ and some slowly varying $\ell_1$, where $\beta$ and $\ell_1$ depend on the distribution of $\xi$. Furthermore, if $\beta = 1/2$ then either $\ell_1(x) = \text{const}$ or $\lim\limits_{x\to\infty} \ell_1(x) = \infty$. Thus, in any case,

$$\frac{x}{c^2(x)} = O(1), \quad x \to \infty. \tag{15}$$

The proof of Theorem 5.1 relies heavily upon the following.

**Lemma 5.3.** *Under the assumption and notation of* Theorem 5.1*,*

$$J(x) := \frac{\int_0^x (\rho(y) - \mathtt{m}^{-1}y)\mathrm{d}y}{xc(x)} \overset{d}{\to} X, \quad x \to \infty. \tag{16}$$

*Proof.* It is known (see Theorem 1b in [7]) that

$$W_x(\cdot) := \frac{\rho(x\cdot) - \mathtt{m}^{-1}(x\cdot)}{c(x)} \Rightarrow Z(\cdot), \quad x \to \infty, \tag{17}$$

in $D[0, \infty)$ in the $M_1$-topology. Since integration is a continuous operator from $D[0, \infty)$ to $D[0, \infty)$, we have

$$\int_0^1 W_x(y)\mathrm{d}y \overset{d}{\to} \int_0^1 Z(y)\mathrm{d}y, \quad x \to \infty,$$

which is equivalent to (16). □

**Remark** When $Z(\cdot)$ is a Brownian motion, the one-dimensional convergence in (16) can be upgraded to the functional limit theorem. Indeed, since $(Z(t))$ is continuous the convergence in (17) is equivalent to the locally uniform convergence. Furthermore, the integration $z(\cdot) \mapsto \int_0^{(\cdot)} z(y)\mathrm{d}y$ is continuous w.r.t. the locally uniform convergence. Hence, by the continuous mapping theorem,

$$\int_0^{(\cdot)} W_x(y)\mathrm{d}y \ \Rightarrow \ \int_0^{(\cdot)} Z(y)\mathrm{d}y, \quad x \to \infty$$

in $D[0,\infty)$.

Lemma 5.4 collects some facts borrowed from [14].

**Lemma 5.4.** (a) $\mathbb{E}(N(x)-M(x))^2 = o(x)$, *as* $x \to \infty$.
(b) *Under the assumption and notation of* Theorem 5.1*,*

$$\frac{\sup\limits_{y\in[0,x]} (\rho(y) - \mathtt{m}^{-1}y)}{c(x)} \ \overset{d}{\to} \ \sup_{t\in[0,1]} Z(t), \quad \text{as} \quad x \to \infty,$$

*and*

$$\frac{\inf\limits_{y\in[0,x]} (\rho(y) - \mathtt{m}^{-1}y)}{c(x)} \ \overset{d}{\to} \ \inf_{t\in[0,1]} Z(t), \quad \text{as} \quad x \to \infty.$$

*Proof of* Theorem 5.1. Applying the Cauchy-Schwarz inequality,

$$\frac{\mathbb{E}\Big( \int_0^x |N(y)-M(y)|\mathrm{d}y\Big)^2}{x^2c^2(x)} \le \frac{\int_0^x \mathbb{E}(N(y)-M(y))^2\mathrm{d}y}{xc^2(x)} = \frac{o(x^2)}{x^2}\frac{x}{c^2(x)},$$

where for the final estimate Lemma 5.4(a) was utilized. In view of (15), the latter expression goes to 0, which implies that

$$\frac{\int_0^x (N(y)-M(y))\mathrm{d}y}{xc(x)} \ \overset{P}{\to} \ 0, \quad x \to \infty. \tag{18}$$

Since

$$\frac{\int_0^x (N(y)-\mathtt{m}^{-1}(y-r(y)))\mathrm{d}y}{xc(x)} = \frac{\int_0^x (N(y)-M(y))\mathrm{d}y}{xc(x)} + \frac{\int_0^x (M(y)-\mathtt{m}^{-1}(y-r(y)))\mathrm{d}y}{xc(x)},$$

we have to prove that the second summand converges in distribution to $X$.

With $\delta \in (0,1)$ such that $y^\delta = o(c(y))$, write for $y > 1$

$$\begin{aligned} F(y) + M(y) - \mathtt{m}^{-1}(y-r(y)) &= \int_0^y (\rho(y-z) - \mathtt{m}^{-1}(y-z))\mathrm{d}F(z) \\ &= \int_0^{y^\delta} \ldots + \int_{y^\delta}^y \ldots \\ &= T_1(y) + T_2(y). \end{aligned}$$

In view of

$$T_1(y) \le (\rho(y) - \mathtt{m}^{-1}y)F(y^\delta) + \mathtt{m}^{-1}y^\delta F(y^\delta) \le (\rho(y) - \mathtt{m}^{-1}y) + \mathtt{m}^{-1}y^\delta$$

we have

$$\frac{\int_0^x T_1(y)\mathrm{d}y}{xc(x)} \le \frac{\int_0^1 T_1(y)\mathrm{d}y}{xc(x)} + \frac{\int_0^x (\rho(y) - \mathtt{m}^{-1}y)\mathrm{d}y}{xc(x)} + \frac{(\delta+1)^{-1}x^\delta}{\mathtt{m}c(x)} \overset{d}{\to} 0 + X + 0 = X,$$

where the last step is justified by Lemma 5.3 and the choice of $\delta$. Further,

$$T_1(y) \ge (\rho(y) - \mathtt{m}^{-1}y) - (\rho(y) - \mathtt{m}^{-1}y)(1 - F(y^\delta)) - (\rho(y) - \rho(y - y^\delta)).$$

Since

$$\frac{\mathbb{E}\int_1^x (\rho(y) - \rho(y - y^\delta))\mathrm{d}y}{xc(x)} \le \frac{\int_1^x \mathbb{E}\rho(y^\delta)\mathrm{d}y}{xc(x)} \le \frac{\mathbb{E}\rho(x^\delta)}{x^\delta}\frac{x^\delta}{c(x)} \to \mathtt{m}^{-1} \cdot 0 = 0,$$

by the elementary renewal theorem and the choice of $\delta$, we conclude that

$$\frac{\int_1^x (\rho(y) - \rho(y - y^\delta))\mathrm{d}y}{xc(x)} \overset{P}{\to} 0.$$

Therefore,

$$\begin{aligned}\frac{\int_0^x T_1(y)\mathrm{d}y}{xc(x)} &\ge \frac{\int_1^x (\rho(y) - \mathtt{m}^{-1}y)\mathrm{d}y}{xc(x)} - \frac{\sup_{0\le y\le x}(\rho(y) - \mathtt{m}^{-1}y)}{c(x)}\frac{\int_1^x (1 - F(y^\delta))\mathrm{d}y}{x} \\ &- \frac{\int_1^x (\rho(y) - \rho(y - y^\delta))\mathrm{d}y}{xc(x)} \\ &\overset{d}{\to} X - 0 - 0 = X,\end{aligned}$$

by Lemma 5.3 and Lemma 5.4 (b).

Finally,

$$\inf_{0\le z\le y}(\rho(z) - \mathtt{m}^{-1}z)(F(y) - F(y^\delta)) \le T_2(y) \le \sup_{0\le z\le y}(\rho(z) - \mathtt{m}^{-1}z)(F(y) - F(y^\delta))$$

entails

$$\frac{\int_0^x T_2(y)\mathrm{d}y}{xc(x)} \le \frac{\int_0^1 T_2(y)\mathrm{d}y}{xc(x)} + \frac{\sup_{0\le z\le x}(\rho(z) - \mathtt{m}^{-1}z)}{c(x)}\frac{\int_1^x (F(y) - F(y^\delta))\mathrm{d}y}{x} \overset{P}{\to} 0,$$

where the last step follows from Lemma 5.4(b) and the trivial fact that the last ratio goes to 0 for any distribution function $F$. Similarly,

$$\frac{\int_0^x T_2(y)\mathrm{d}y}{xc(x)} \ge \frac{\int_0^1 T_2(y)\mathrm{d}y}{xc(x)} + \frac{\inf_{0\le z\le x}(\rho(z) - \mathtt{m}^{-1}z)}{c(x)}\frac{\int_1^x (F(y) - F(y^\delta))\mathrm{d}y}{x} \overset{P}{\to} 0,$$

Putting the pieces together completes the proof. □

## 5.2 Proof of Theorem 3.2 and Theorem 3.1

*Proof of Theorem 3.2.* We shall make use of the Poissonized version of the occupancy model with random frequencies $(P_k)$, in which balls are thrown in boxes at epochs of a unit rate Poisson process $(\pi_t)_{t\geq 0}$. For simplicity we use notation $V(t) = \log T_{\pi_t}$.

Set

$$\rho^*(x) := \inf\{k \in \mathbb{N} : W_1 \dots W_k < e^{-x}\}, \quad x \geq 0,$$

and

$$\begin{aligned} N^*(x) &:= \#\{k \in \mathbb{N} : P_k \geq e^{-x}\} \\ &= \#\{k \in \mathbb{N} : W_1 \cdots W_{k-1}(1 - W_k) \geq e^{-x}\}, \quad x \geq 0. \end{aligned}$$

First of all, we need a refined large deviation result for $(\pi_t)$ itself: for $t > 1$,

$$\mathbb{P}\{\pi_t \leq (1 - \varepsilon_t)t\} \leq \exp(-t(\varepsilon_t + \log(1 - \varepsilon_t)(1 - \varepsilon_t))) =: q(t), \tag{19}$$

where $\varepsilon_t := t^{-\beta}$, for any $\beta \in (0, 1/2)$. Note that $\lim_{t\to\infty} q(t) = 0$ with $(-\log q(t)) \sim t^{1-2\beta}$. Inequality (19) is the Chernoff bound for the Poisson distribution and follows in a standard way by first applying Markov's inequality to $e^{-s\pi_t}$ and then minimizing the right-hand side over $s$.

For $j = 1, 2$, set

$$f_j(t) := \mathbb{E}(\log^+ \pi_t)^j = e^{-t} \sum_{k\geq 2} \log^j k(t^k/k!), \quad t \geq 0.$$

These functions are nondecreasing and differentiable with $f_j(0) = 0$ and

$$f_j'(0) = 0. \tag{20}$$

Let us prove that

$$\lim_{t\to\infty} (f_1(t) - \log t) = 0 \tag{21}$$

and

$$\lim_{t\to\infty} h(t) = 0, \tag{22}$$

where $h(t) := \mathrm{Var}(\log^+ \pi_t)$. To this end, write

$$f_1(t) - \log t \leq \mathbb{E}\log(\pi_t + 1) - \log t \leq \log(t + 1) - \log t \leq t^{-1}, \tag{23}$$

where at the second step Jensen's inequality has been utilized. Similarly,

$$\begin{aligned} f_2(t) - \log^2 t &\leq \mathbb{E}\log^2(\pi_t + 1) - \log^2 t \\ &\leq \log^2(t + 1) - \log^2 t \\ &\leq 2t^{-1}\log(t + 1). \end{aligned} \tag{24}$$

Note that we actually work on the set $\{\pi_t \geq 2\}$ and that the function $t \mapsto \log^2(1 + t)$ is concave for $t \geq 2$.

Furthermore, for large enough $t$, and $\varepsilon_t$ as defined above,

$$\begin{aligned} f_1(t) - \log t &\geq \mathbb{E}(\log^+ \pi_t - \log t) 1_{\{\pi_t > (1-\varepsilon_t)t\}} - \log t \mathbb{P}\{\pi_t \leq (1-\varepsilon_t)t\} \\ &\geq \log(1-\varepsilon_t)\mathbb{P}\{\pi_t > (1-\varepsilon_t)t\} - q(t)\log t =: p(t). \end{aligned}$$

and the last expression goes to zero (with rate $t^{-\beta}$), as $t \to \infty$. Combining this inequality with (23) proves (21). Note also that

$$\begin{aligned} f_1^2(t) &= \log^2 t + 2\log t(f_1(t) - \log t) + (f_1(t) - \log t)^2 \\ &\geq \log^2 t + 2p(t)\log t. \end{aligned} \tag{25}$$

Hence

$$h(t) = f_2(t) - f_1^2(t) \overset{(24),(25)}{\leq} 2(t^{-1}\log(t+1) - p(t)\log t) = O(\log t/t^{\beta}),$$

which proves (22)[1].

The basic observations for the subsequent work are given and proved next:

$$\begin{aligned} \mathbb{E}(V(t)|(P_k)) &= \sum_{j\geq 1} f_1(tP_j) \\ &= \int_1^\infty f_1(t/x)\mathrm{d}N^*(\log x) \\ &= \int_0^{\log t} (\log t - x)\mathrm{d}N^*(x) + O_P(\log t) \qquad (26) \\ &= \int_0^{\log t} N^*(x)\mathrm{d}x + O_P(\log t) \end{aligned}$$

and

$$\begin{aligned} \operatorname{Var}(V(t)|(P_k)) &= \sum_{j\geq 1} h(tP_j) \\ &= O_P(\log t), \end{aligned} \tag{27}$$

where $O_P(\log t)$ means that $O_P(\log t)/\log t$ is bounded in probability.

The a.s. finiteness of the conditional expectation (and even its integrability) can be justified as follows:

$$\mathbb{E}\log T_n \leq (\log^+ n)\mathbb{E}K_n \leq n\log^+ n.$$

Hence $\mathbb{E}V(t) \leq \mathbb{E}\pi_t \log^+ \pi_t < \infty$. The integrability of the conditional variance can be checked similarly.

Since $N^*(\log y) \leq \rho^*(\log y)$, and $\rho^*(\log y) = O_P(\log y)$ we conclude that

$$N^*(\log y) = O_P(\log y). \tag{28}$$

[1] Alternatively, both (21) and (22) can be deduced from Theorem 4 in [23]. To keep the paper self-contained we prefer to give an elementary real-analytic argument.

Using this and (21) gives

$$\int_1^t f_1(t/x)\mathrm{d}N^*(\log x) = \int_0^{\log t} (\log t - x)\mathrm{d}N^*(x) + O_P(\log t).$$

In fact, only boundedness of $f_1(t) - \log t$ was used. Further,

$$\begin{aligned}\int_t^\infty f_1(t/x)\mathrm{d}N^*(\log x) &= -f_1(1)N^*(\log t) + \int_0^1 N^*(\log t - \log x)f_1'(x)\mathrm{d}x \\ &\overset{(28)}{\leq} O_P(\log t) + \rho^*(\log t)f_1(1) \\ &+ \int_0^1 (\rho^*(\log t - \log x) - \rho^*(\log t))f_1'(x)\mathrm{d}x \\ &= O_P(\log t),\end{aligned}$$

since by the well-known bound for the renewal function

$$\mathbb{E}\int_0^1 (\rho^*(\log t - \log x) - \rho^*(\log t))f_1'(x)\mathrm{d}x \leq \int_0^1 (C_1|\log x| + C_2)f_1'(x)\mathrm{d}x \overset{(20)}{<} \infty,$$

where $C_1$ and $C_2$ are positive constants. Thus we have proved (26). The proof of (27) follows the same pattern, the only minor difference being that now we use inequality

$$\int_t^\infty h(t/x)\mathrm{d}N^*(\log x) \leq \int_t^\infty f_2(t/x)\mathrm{d}N^*(\log x)$$

and (20) for $f_2$.

Throughout the rest of the proof we apply results of Subsection 5.1 to the vector $(\xi, \eta) := (|\log W|, |\log(1-W)|)$. With this specific choice the quantities $\rho(x)$ and $N(x)$ defined in Subsection 5.1 turn into $\rho^*(x)$ and $N^*(x)$.

Let $(X(t))_{t\geq 0}$ be a Lévy process with $\log \mathbb{E}e^{izX(1)} = \psi(z)$, $z \in \mathbb{R}$. Then

$$\log \mathbb{E}\exp\left(iz\int_0^1 X(t)\mathrm{d}t\right) = \int_0^1 \psi(zs)\mathrm{d}s, \tag{29}$$

which follows from a Riemann approximation to the integral.

Assume that the assumptions of Theorem 3.2 hold which implies that the assumption of Theorem 5.1 (with $\rho$ replaced by $\rho^*$) holds. By scaling $Z$ and $c(x)$, if necessary, we can assume that $Z$ has the standard normal distribution under the assumptions of parts (a) and (b) of Theorem 3.2 and that $Z$ has a stable law with characteristic function (8) under (7). Then (29) implies that $X \overset{d}{=} 3^{-1/2}Z \overset{d}{=} \mathcal{N}(0, 1/3)$ in the first case, and that $X \overset{d}{=} (\alpha+1)^{-1/\alpha}Z$ in the second case.

By Theorem 5.1,

$$\frac{\int_0^{\log t}(N^*(y) - \mu^{-1}(y - r^*(y)))\mathrm{d}y}{c(\log t)\log t} \overset{d}{\to} X, \quad t \to \infty,$$

where $r^*(y) := \int_0^y \mathbb{P}\{|\log(1-W)| > z\}\mathrm{d}z$. Since $\lim_{t\to\infty} c(t) = \infty$, using (26) yields

$$\frac{\mathbb{E}(V(t)|(P_k)) - \mu^{-1}\Big(2^{-1}\log^2 t - \int_0^{\log t} r^*(y)\mathrm{d}y\Big)}{c(\log t)\log t} \overset{d}{\to} X, \quad t\to\infty,$$

and hence

$$\frac{V(t) - \mu^{-1}\Big(2^{-1}\log^2 t - \int_0^{\log t} r^*(y)\mathrm{d}y\Big)}{c(\log t)\log t} \overset{d}{\to} X, \quad t\to\infty,$$

by virtue of (27) and Chebyshev's inequality.

Now we have to de-Poissonize, i.e., to pass from the Poissonized occupancy model to the fixed-$n$ model. This is simple as $(\log T_n)$ is a nondecreasing sequence. Set

$$b(t) := \mu^{-1}\bigg(2^{-1}\log^2 t - \int_0^{\log t} \mathbb{P}\{|\log(1-W)| > y\}\mathrm{d}y\bigg) \ \text{ and } \ a(t) := c(\log t)\log t.$$

Recall that we take a properly adjusted $c(x)$. Since $a(t)$ grows faster than the logarithm, we have

$$\lim_{t\to\infty}\frac{b(t) - b(\lfloor t(1\pm\varepsilon)\rfloor)}{a(t)} = 0,$$

for every $\varepsilon > 0$. This together with slow variation of $a(t)$ give

$$X_\pm(t) := \frac{V(t) - b(\lfloor t(1\pm\varepsilon)\rfloor)}{a(\lfloor t(1\pm\varepsilon)\rfloor)} \overset{d}{\to} X.$$

By the monotonicity of $(\log T_n)$, we have

$$\begin{aligned} X_+(t) &= X_+(t)1_{D_t} + X_+(t)1_{(D_t)^c} \\ &\leq \frac{V_{\lfloor(1+\varepsilon)t\rfloor} - b(\lfloor t(1+\varepsilon)\rfloor)}{a(\lfloor t(1+\varepsilon)\rfloor)}1_{D_t} + X_+(t)1_{(D_t)^c}, \end{aligned}$$

where $D_t := \big\{\pi_t \in [\lfloor(1-\varepsilon)t\rfloor, \lfloor(1+\varepsilon)t\rfloor]\big\}$. Since $\mathbb{P}(D_t) \to 1$, hence $X_+(t)1_{(D_t)^c} \overset{P}{\to} 0$, we conclude that

$$\mathbb{P}\{X > x\} \leq \liminf_{n\to\infty}\mathbb{P}\bigg\{\frac{\log T_n - b(n)}{a(n)} > x\bigg\},$$

for all $x \in \mathbb{R}$. To prove the converse inequality for the upper bound one can proceed in a similar manner.

It remains to set $b_n = b(n)$, and $a_n = (\alpha+1)^{-1/\alpha}a(n)$ if the assumption of part (c) holds, and $a_n = 3^{-1/2}a(n)$ if the assumptions of parts (a) and (b) hold. The fact that the so-defined $a_n$ and $b_n$ are of the form as stated in Theorem 3.2 follows from considerations above and from, for instance, Proposition 27 in [28]. The proof of Theorem 3.2 is complete.
*Proof of Theorem 3.1.* By Theorem 3.2, $(\log T_n - b_n)/a_n$, with case-dependent $a_n$ and $b_n$ defined in Theorem 3.1, weakly converges. In particular, we know that $\log^{3/2} n = O(a_n)$. It remains to apply Lemma 4.2 and Markov's inequality. The proof of Theorem 3.1 is complete.

# 6 Appendix

The following lemma is a simple consequence of Proposition 3 in [10].

**Lemma 6.1.** *Assume that the sequence $a_n$ satisfies the following recurrence relation*

$$a_0 = 0, \quad a_n = b_n + \sum_{k=0}^{n} q(n,k) a_{n-k}, \quad n \in \mathbb{N}.$$

*Then*

*(i) if $b_n = O(1)$ then $a_n = O(\log n)$, as $n \to \infty$,*

*(ii) if $b_n = O(\log n)$ then $a_n = O(\log^2 n)$, as $n \to \infty$.*

The next lemma verifies (14) which is a key ingredient of the proof of Lemma 4.1.

**Lemma 6.2.** *Relation* (14) *holds provided the density $f$ of $W$ satisfies any of the following two conditions:*

(i) *condition* (9) *holds for some $\alpha \in [0,1)$.*

(ii) *there exist $\delta_1 \geq 0$ and $\delta_2 \geq 0$ such that $f$ is nonincreasing on $(0,\delta_1)$, bounded on $[\delta_1, 1-\delta_2]$ and nondecreasing on $(1-\delta_2, 1)$.*

*Proof.* We start with easier part (i). We have

$$\begin{aligned}
k \sum_{r=1}^{\lfloor n/k \rfloor - 1} \mathbb{P}\{A_n = rk\} &= \frac{k}{1-\mathbb{E}W^n} \sum_{r=1}^{\lfloor n/k \rfloor - 1} \binom{n}{rk} \int_0^1 x^{n-rk}(1-x)^{rk} f(x)\mathrm{d}x \\
&\leq \text{const}\, \frac{k}{1-\mathbb{E}W^n} \sum_{r=1}^{\lfloor n/k \rfloor - 1} \binom{n}{rk} \int_0^1 x^{n-rk-\alpha}(1-x)^{rk-\alpha}\mathrm{d}x \\
&= \text{const}\, \frac{k}{1-\mathbb{E}W^n} \sum_{r=1}^{\lfloor n/k \rfloor - 1} \frac{\Gamma(n+1)\Gamma(n-rk-\alpha+1)\Gamma(rk-\alpha+1)}{\Gamma(n-2\alpha+2)\Gamma(n-rk+1)\Gamma(rk+1)} \\
&\leq \text{const}\, \frac{1}{1-\mathbb{E}W^n} \frac{k}{n^{1-2\alpha}} \sum_{r=1}^{\lfloor n/k \rfloor - 1} ((n-rk)rk)^{-\alpha} \\
&\leq \text{const}\, \frac{1}{1-\mathbb{E}W^n} \frac{k^{1-2\alpha}}{n^{1-2\alpha}} \sum_{r=1}^{\lfloor n/k \rfloor - 1} ((\lfloor n/k \rfloor - r)r)^{-\alpha} = O(1),
\end{aligned}$$

The fourth line is a consequence of the inequality given in [1], formula (6.1.47): for $c, d > -1$ there exists $M_{c,d} > 0$ such that for all $n \in \mathbb{N}$

$$\left| \frac{\Gamma(n+c)}{\Gamma(n+d)} - n^{c-d} \right| \leq M_{c,d} n^{c-d-1}.$$

The equality in the last line follows from the estimate $\sum_{j=1}^{m-1}((m-j)j)^{-\alpha} \leq \text{const}\ m^{1-2\alpha}$ which holds for $\alpha < 1$ and $m \in \mathbb{N}$. The proof of part (i) is complete.

Passing to part (ii) we can write

$$f(x) = \mathbb{P}\{W \leq \delta_1\}f_1(x) + \mathbb{P}\{\delta_1 < W \leq 1-\delta_2\}f_2(x) + \mathbb{P}\{W > 1-\delta_2\}f_3(x), \tag{30}$$

where $f_1$, $f_2$ and $f_3$ are some densities such that $f_1$ is nonincreasing on $(0,1)$, $f_3$ is nondecreasing on $(0,1)$ and $f_2$ is bounded on $(0,1)$. It is known (see [22]) that if a random variable $X$ with support $[0,1]$ has a nonincreasing (nondecreasing) density $h$ then there exists a distribution function $G$ such that $h(x) = \int_x^1 \frac{\mathrm{d}G(y)}{y}$ (resp. $h(x) = \int_{1-x}^1 \frac{\mathrm{d}G(y)}{y}$). Using this observation (30) can be rewritten as follows

$$\begin{aligned} f(x) &= \mathbb{P}\{W \leq \delta_1\}\int_0^1 \frac{1_{\{x\in[0,y]\}}}{y}\mathrm{d}G_1(y) + \mathbb{P}\{\delta_1 < W \leq 1-\delta_2\}f_2(x) \\ &+ \mathbb{P}\{W > 1-\delta_2\}\int_0^1 \frac{1_{\{x\in[1-y,1]\}}}{y}\mathrm{d}G_2(y), \end{aligned}$$

where $G_1, G_2$ are some distribution functions concentrated on $[0,\delta_1]$ and $[1-\delta_2,1]$, respectively.

The last formula can be seen as a representation of $f$ as a convex linear combination of the densities of three types: $g_\varepsilon(x) = \varepsilon^{-1}1_{\{x\in[0,\varepsilon]\}}$, $h_\varepsilon(x) = \varepsilon^{-1}1_{\{x\in[1-\varepsilon,1]\}}$ and bounded densities. Thus to prove (ii) it is enough to show that relation (14) holds for densities of these types uniformly in $\varepsilon \in (0,1)$. The validity of (14) for bounded densities follows from part (i) of the lemma (take $\alpha = 0$). We only check (14) for $g_\varepsilon$, as the argument is symmetric for $h_\varepsilon$. We have

$$\mathbb{P}\{A_n = k\} = \binom{n}{k}\varepsilon^{-1}\int_0^\varepsilon p^k(1-p)^{n-k}\mathrm{d}p = \frac{1}{(n+1)\varepsilon}I_\varepsilon(k+1, n-k+1),$$

where $I_\varepsilon(k+1, n-k+1)$ is the normalized truncated beta-function (see formula (6.6.2) in [1]). Using formulae (6.6.5) and (6.6.4) of the same reference we obtain

$$\mathbb{P}\{A_n = k\} = \frac{1}{n+1}I_\varepsilon(k, n-k+1) + \frac{1-\varepsilon}{(n+1)\varepsilon}\mathbb{P}\{B \geq k+1\} \leq \frac{1}{n+1} + \frac{1}{(n+1)\varepsilon}\mathbb{P}\{B \geq k+1\}$$

where a random variable $B$ has the binomial distribution with parameters $(n,\varepsilon)$. This yields

$$
\begin{aligned}
k\sum_{r=1}^{\lfloor n/k\rfloor-1}\mathbb{P}\{A_n=rk\} &\le k\sum_{r=1}^{\lfloor n/k\rfloor-1}\Big(\frac{1}{n+1}+\frac{1}{(n+1)\varepsilon}\mathbb{P}\{B\ge rk+1\}\Big)\\
&\le 1+\frac{k}{(n+1)\varepsilon}\sum_{r=1}^{\lfloor n/k\rfloor-1}\mathbb{P}\{B\ge rk+1\}\\
&= 1+\frac{k}{(n+1)\varepsilon}\sum_{r=1}^{\lfloor n/k\rfloor-1}\sum_{j=rk+1}^{n}\mathbb{P}\{B=j\}\\
&\le 1+\frac{k}{(n+1)\varepsilon}\sum_{j=1}^{n}\sum_{r=1}^{\lfloor j/k\rfloor}\mathbb{P}\{B=j\}\\
&\le 1+\frac{1}{(n+1)\varepsilon}\sum_{j=1}^{n}j\mathbb{P}\{B=j\}\le 2.
\end{aligned}
$$

The proof of part (ii) is complete. □

**Lemma 6.3.** *Let* $(b_n(k))_{n\in\mathbb{N},1\le k\le n}$, $(c_n)_{n\in\mathbb{N}}$ *and* $(d_n)_{n\in\mathbb{N}}$ *be nonnegative arrays. Let* $(a_n(k))_{n\in\mathbb{N}_0,\,k\in\mathbb{N}}$ *and* $(a'_n)_{n\in\mathbb{N}_0}$ *be defined recursively via*

$$
\begin{gathered}
a_0(k)=a_1(k)=\ldots=a_{k-1}(k)=0,\quad k\in\mathbb{N};\\
a_n(k)=b_n(k)+\sum_{i=k}^{n-1}p_{n,i}a_i(k),\quad k\le n,\ k\in\mathbb{N};
\end{gathered}
$$

*and*

$$a'_0=0,\ \ a'_n=d_n+\sum_{i=0}^{n-1}p_{n,i}a'_i,\quad n\in\mathbb{N},$$

*respectively, where* $(p_{n,k})_{0\le k\le n-1}$ *is a probability distribution, for every fixed* $n\in\mathbb{N}$.

*If*

$$c_kb_n(k)\le d_n,\quad n\in\mathbb{N},\ \ k\le n,\ k\in\mathbb{N}, \tag{31}$$

*then*

$$c_ka_n(k)\le a'_n,\quad n\in\mathbb{N},\ \ k\le n,\ k\in\mathbb{N}. \tag{32}$$

*Proof.* We shall prove the lemma by induction on $n$. The base of induction is straightforward. Assume that (32) holds for all positive integer $n\le N$ and $k\le n$. We have to prove (32) for $n=N+1$ and $k\le N+1$, $k\in\mathbb{N}$. Assume first that $k\le N$, then

$$
\begin{aligned}
c_ka_{N+1}(k) &= c_kb_{N+1}(k)+\sum_{i=k}^{N}p_{N+1,\,i}c_ka_i(k)\overset{(31)}{\le} d_{N+1}+\sum_{i=k}^{N}p_{N+1,\,i}c_ka_i(k)\\
&\overset{\text{induction}}{\le} d_{N+1}+\sum_{i=k}^{N}p_{N+1,\,i}a'_i\le d_{N+1}+\sum_{i=0}^{N}p_{N+1,\,i}a'_i=a'_{N+1}.
\end{aligned}
$$

For $k = N + 1$ we have

$$c_{N+1}a_{N+1}(N+1) = c_{N+1}b_{N+1}(N+1) \leq d_{N+1} \leq a'_{N+1}.$$

The proof is complete. □

**Acknowledgement** The authors are indebted to a referee for thoughtful comments which have led to a number of improvements. The work of the second author was partially supported by the Department of Mathematics of the Utrecht University and the Dutch stochastics cluster STAR.

# References

[1] Abramowitz, M. and Stegun, I. (1964). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover.

[2] Apostol T. M. (1976). Introduction to Analytic Number Theory. New York: Springer-Verlag.

[3] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic combinatorial structures*, European Mathematical Society.

[4] Arratia, R. and Tavaré, S. (1992). Limit theorems for combinatorial structures via discrete process approximations. *Random Struct. Algorithms* **3**, 321–345.

[5] Babu, G. J. and Manstavičius, E. (2002). Limit processes with independent increments for the Ewens sampling formula. *Ann. Inst. Stat. Math.* **54**, 607–620.

[6] Betz, V., Ueltschi, D. and Velenik, Y. (2011). Random permutations with cycle weights. *Ann. Appl. Prob.* **21**, 312–331.

[7] Bingham, N. H. (1973). Maxima of sums of random variables and suprema of stable processes. *Z. Wahrsch. verw. Gebiete.* **26**, 273– 296.

[8] Diaconis, P. (1988) *Group representations in probability and statistics*, IMS Lecture Notes–Monograph Series, Volume 11 Institute of Mathematical Statistics, Hayward, CA.

[9] Erdös, P. and Turán, P. (1967). On some problems of statistical group theory III. *Acta. Math. Acad. Sci. Hungar.* **18**, 309–320.

[10] Gnedin, A. (2004). The Bernoulli sieve. *Bernoulli* **10**, 79–96.

[11] Gnedin, A. (2004). Three sampling formulas. *Combinatorics, Probability and Computing*, **13**, 185–193.

[12] Gnedin, A. (2011). Coherent random permutations with biased record statistics. *Discrete Mathematics* **311**, 80–91.

[13] Gnedin, A., Haulk, C. and Pitman, J. (2010). Characterizations of exchangeable partitions and random discrete distributions by deletion properties. *London Mathematical Society Lecture Notes Series* **378**, 264–298.

[14] Gnedin, A., Iksanov, A. and Marynych, A. (2010). Limit theorems for the number of occupied boxes in the Bernoulli sieve. *Theory of Stochastic Processes* **16(32)**, 44–57.

[15] Gnedin, A., Iksanov, A. and Marynych, A. (2010). The Bernoulli sieve: an overview. *Discr. Math. Theoret. Comput. Sci.* Proceedings Series, **AM**, 329–342.

[16] Gnedin, A., Iksanov, A., Negadajlov, P. and Roesler, U. (2009). The Bernoulli sieve revisited. *Ann. Appl. Prob.* **19**, 1634–1655.

[17] Gnedin, A., Iksanov, A. and Roesler, U. (2008). Small parts in the Bernoulli sieve. *Discr. Math. Theoret. Comput. Sci.* Proceedings Series, **AI**, 239–246.

[18] Gnedin, A. and Olshanski, G. (2006). Coherent permutations with descent statistic and the boundary problem for the graph of zigzag diagrams. *Intern. Math. Res. Not.* Art. 51968, 1–39.

[19] Gnedin, A. and Pitman, J. (2005). Regenerative composition structures. *Ann. Probab.* **33**, 445–479.

[20] Iksanov, A. (2012+). On the number of empty boxes in the Bernoulli sieve I. *Stochastics*, to appear.

[21] Iksanov, A. (2012+). On the number of empty boxes in the Bernoulli sieve II. *Stoch. Proc. Appl.*, to appear.

[22] Lukacs, E. (1970). *Characteristic functions.* London: Charles Griffin & Company.

[23] Jacquet, P. and Szpankowski, W. (1999). Entropy computations via analytic de-Poissonization. *IEEE Trans. Inform. Theory.* **45**, 1072–1081.

[24] DeLaurentis, J. M. and Pittel, B. G. (1985). Random permutations and Brownian motion. *Pacific J. Math.* **119**, 287–301.

[25] Manstavičius, E. (2009). An analytic method in probabilistic combinatorics. *Osaka J. Math.* **46**, 273-290.

[26] Medvedev, Yu. I. (1977). Separable statistics in a polynomial scheme. II. *Theory Probab. Appl.* **22**, 607–615.

[27] Mirakhmedov, Sh. A. (1989). Randomized decomposable statistics in a generalized allocation scheme over a countable set of cells. *Diskretnaya Matematika.* **1**, 46–62.

[28] Negadailov, P. (2010). Limit theorems for random recurrences and renewal-type processes. PhD thesis, Utrecht University. Available at http://igitur-archive.library.uu.nl/dissertations/

[29] Pitman, J. (2006). *Combinatorial Stochastic Processes*, Berlin: Springer-Verlag.